Towards Condition-Invariant Sequence-Based Route Recognition

Michael Milford
School of Engineering Systems
Queensland University of Technology
michael.milford@qut.edu.au

Abstract

In this paper we present a novel algorithm for localization during navigation that performs matching over local image sequences. Instead of calculating the single location most likely to correspond to a current visual scene, the approach finds candidate matching locations within every section (subroute) of all learned routes. Through this approach, we reduce the demands upon the image processing front-end, requiring it to only be able to correctly pick the best matching image from within a short local image sequence, rather than globally. We applied this algorithm to a challenging downhill mountainbiking visual dataset where there was significant perceptual or environment change between repeated traverses of the environment, and compared performance to applying the feature-based algorithm FAB-MAP. The results demonstrate the potential for localization using visual sequences, even when there are no visual features that can be reliably detected.

1 Introduction

Visual recognition under any illumination and environmental conditions is a holy grail for robotics and computer vision, and is a task far beyond current state of the art algorithms in the sub domain of visual navigation. Current visual SLAM systems are predominantly tailored towards creating large and accurate maps of environments, sometimes in an online manner. The majority of visual SLAM systems developed thus far have been based around high quality stereo [1] or panoramic visual sensor data [2], although some researchers have focused on single camera systems [3, 4]. Relatively little work has addressed reliable mapping when environmental conditions change significantly, such as day turning to night, sunshine turning to torrential rain, or summer turning to winter [5, 6]. This challenge involves the problem of finding features in an image that are consistent across extreme perceptual change, as illustrated in Figure 1. Consequently robot and personal navigation systems rely primarily on GPS, laser range finders or external beacons, voiding the many advantages of visual sensors such as cheap cost, small size, passive sensing, low power usage, and indoor usage.

In this paper we present a new algorithm that performs matching during navigation over local image sequences. Instead of calculating global matches on a frame by frame basis, the algorithm finds many candidate matches within every local section or subroute of all previously learned routes. Our appearance-based approach is intended for application in environments where robots tend to follow

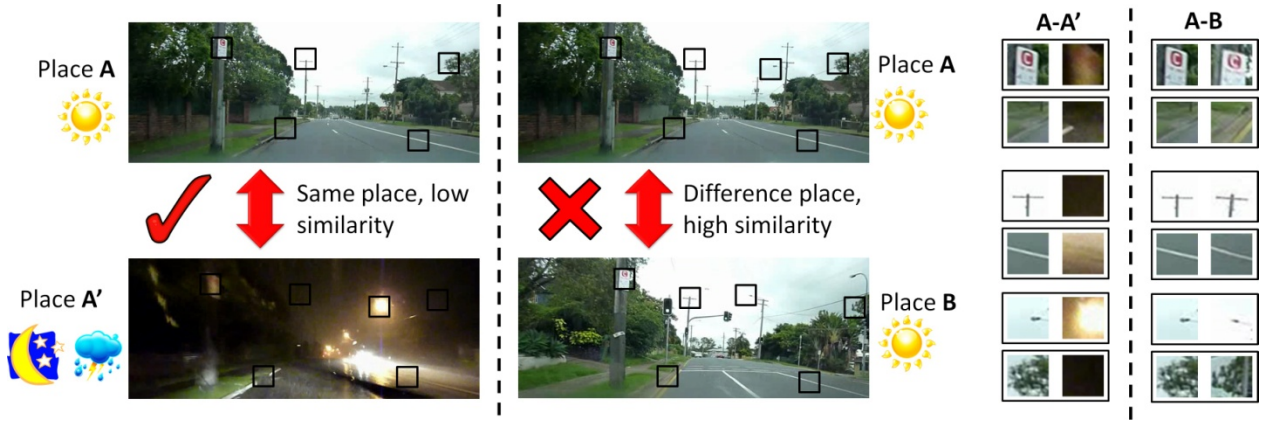

Figure 1 – The problem of perceptual change. Place A and B are different places, yet appear much more similar than Place A does to itself during a rainy night (A'), both on a whole image and individual feature basis. Our approach uses local matching within sequences to successfully match A to A', but not A to B.

somewhat repeatable paths, such as indoor behaviour-based robots [7, 8] or perhaps in future applications on road-based vehicles [9].

The approach alleviates the need for a robust feature matching front end such as Scale-Invariant Feature Transforms (SIFT) [10] and Speeded Up Robust Features (SURF) [11]. We demonstrate the performance of the algorithm using videos taken from onboard a mountain bike during two different runs down a mountain bike trail at high speed. The videos are low quality, taken using different mounting points, and the runs vary significantly in environmental conditions and the path the rider takes. Furthermore, the combination of the high speed, jerky nature of the ride, lighting conditions and video quality mean features are few and far between. For a comparison, we also run FAB-MAP 1.0 on the dataset. Finally, we present preliminary results from a much more challenging dataset, mapping from sunny daytime to dark night-time in the middle of a tropical thunderstorm.

The paper proceeds as follows. In Section 2 we provide some background on vision-based mapping and navigation, and motivate the perspective that has driven this work. Section 3 presents the sequence matching algorithm, while Section 4 details the experimental set up and image pre-processing. The results are presented in Section 5 for both the mountain bike dataset and the day-night dataset (preliminary results in the latter case). Finally the paper concludes with discussion and future work in Section 6.

2 Vision-Based Mapping and Navigation

Visual recognition of places and place sequences in the field of robotic navigation has advanced rapidly in recent years, with several techniques [1, 3, 12-15] now being competitive with laser range-based techniques. Much of this work has occurred under the banner of Simultaneous Localization And Mapping (SLAM) research [16], which focuses on enabling robots to create maps of the environment, while simultaneously using those maps to work out where they are (localize). Vision has been successfully used as the primary sensory modality for learning and recognizing places along journeys of up to 1000 km [12]. Environment size and accuracy have been the dominant benchmarks thus far. Achievements towards this end include MonoSLAM [17], FrameSLAM [1], Fab-MAP, which has mapped a 1000 km car journey [12], and a 142 km car journey using stereo data and bundle adjustment [18].

The problem of visual place recognition under varying lighting, weather and seasonal conditions has received far less attention, with most attempts to address the problem focused on relatively minor issues of perceptual change [6, 19]. Many place recognition techniques rely on feature-finding algorithms such as SIFT [10] and SURF [11] which, despite their impressive rotation and scale invariant properties, are inherently unsuitable when dealing with extreme perceptual change [20]. For example, a vision-only mapping system based on SIFT or SURF will degrade in performance when summer turns to winter and the absence of leaves on trees changes the appearance of the landscape. Some extensions such as C-SIFT [21], OpponentSIFT [22] and RGB-SIFT go some way towards addressing changing environmental conditions. However, when perceptual change becomes extreme, such as day turning to night in a mostly unilluminated environment, these systems fail completely. There have been attempts to increase the robustness of vision by performing sensor fusion with lasers at the level of landmarks [23] or dense depth maps [24]. However, these approaches require sensor-sensor calibration to ensure features are represented in a common reference frame, and are unsuitable in unstructured environments or changing conditions where geometry or features are not reliable cues. Improved sensing hardware and high dynamic range techniques can go some way towards addressing illumination change, but there are limitations to what can be achieved in cheap sensors and on rapidly moving platforms. One promising step [25] towards reducing the reliance on consistent feature detection uses the self-similarity of objects within a frame or video sequence to robustly recognize objects under varying conditions. However, even this approach fails when the environmental change is so great that objects exhibit no consistent self-similarity, as in the transition from day to night in which an off street light becomes a bright flare, or a textured tree becomes a dark uniform blob.

An alternative approach to vision-based mapping has been pursued by [26] in development of the RatSLAM system, a robot SLAM and navigation system based on computational models of the rat hippocampus [7, 9, 26]. RatSLAM has been demonstrated in a number of significant experiments, including mapping of the longest path by a visual SLAM algorithm at the time [9], and a long term delivery robot experiment [7]. The significance of the RatSLAM results is that by localizing over sequences, rather than individual places, the requirement for perfect data association or feature detection is removed. Indeed, significant experimental results were achieved using only low resolution visual sensory input (typically around 1000 pixels [27]) and lightweight pre-processing techniques such as image intensity profiles, without performing traditional feature extraction. More recently, RatSLAM has been combined with FAB-MAP to demonstrate the potential for SLAM performance over multiple times of day [6]. However, recall rates at high levels of precision during that experiment were only around 10 to 15%, and the datasets were less challenging with an overall relatively low level of perceptual change.

3 Sequence Matching for Visual Navigation

In this section we describe the local best match and sequence recognition techniques. The specific frame comparison technique used is not integral to the core algorithm, and is consequently presented in the experimental setup section.

3.1 Template Learning

New visual templates are learned at a fixed rate of V_{av} templates per frame processed. For the experiments described in this paper, we found a value of $V_{av} = 0.5$ gave good results – the only tradeoff being compute speed. Due to significant image redundancy, we only processed one in every four original video frames, so a template was learned every 8 original video frames.

3.2 Local Best Match

We frame the image recognition problem not as one of finding the single template that best matches the current image (global best-match), but rather as one of finding all the templates within local neighborhoods that are the best match for the current image (local best-match). Towards this aim, we apply a local contrast enhancement (analogous to a 1D version of patch normalization) process to each element i in the image difference vector $\mathbf{D}$ (the vector of differences between the current image and all learnt images) to produce a contrast enhanced image difference vector $\hat{\mathbf{D}}$:

$$\hat{D}_i = \frac{D_i - \overline{D_l}}{\sigma_l} \tag{1}$$

where $\overline{D_l}$ is the local mean and σ_l is the local standard deviation, in a range of R_{window} templates around template i. Figure 2 shows a schematic of the local contrast enhancement process operating on a number of $\hat{\mathbf{D}}$ vectors calculated at different times.

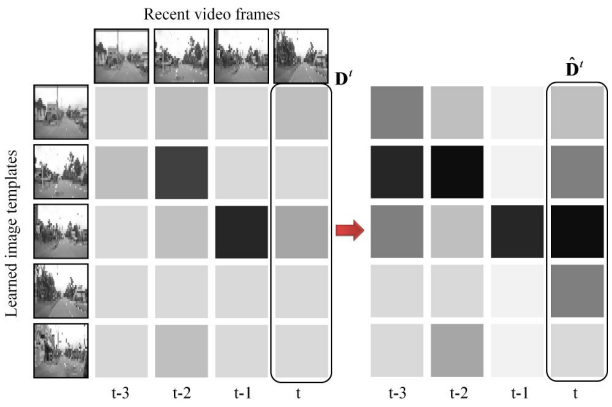

Figure 2: Contrast enhancement of the original image difference vectors increases the number of strongly matching templates. Darker shading = smaller image difference = stronger match.

3.3 Localized Sequence Recognition

Localized template matching produces a number of candidate template matches at each time step. To recognize familiar place sequences, a search is performed through the space $\mathbf{M}$ of recent image difference vectors:

$$\mathbf{M} = \left[\hat{\mathbf{D}}^{T-d_S}, \hat{\mathbf{D}}^{T-d_S+1} \ldots \hat{\mathbf{D}}^{T} \right] \tag{2}$$

where d_s determines how far back in time the search goes, and T is the current time. We make the assumption that velocities (in terms of templates learned per second) on repeated traverses of a path are approximately repeatable within an allowable range (see Table 1). The search projects a number of trajectories representing different possible velocities, starting from each element in $\hat{\mathbf{D}}^{T-d_s}$. Figure 3 shows trajectory lines for various speeds originating from a single template – for clarity reasons searches originating from the other templates are not shown. A score S is calculated for each trajectory line based on the difference values the line passes through in travelling from time T-d_s to the current time T:

$$S = \sum_{t=T-d_S}^{T} D_k^t \tag{3}$$

where k is the particular difference value the trajectory passes through at time t:

$$k = s + V\left(d_s - T + t\right) \tag{4}$$

where s is the template number the trajectory originated in, and V is the trajectory velocity which is varied between V_{min} and V_{max} in steps of V_{step}. Trajectory searches are performed from every template except for recently learned templates within R_{recent} of the current template, in order to avoid matching the current trajectory.

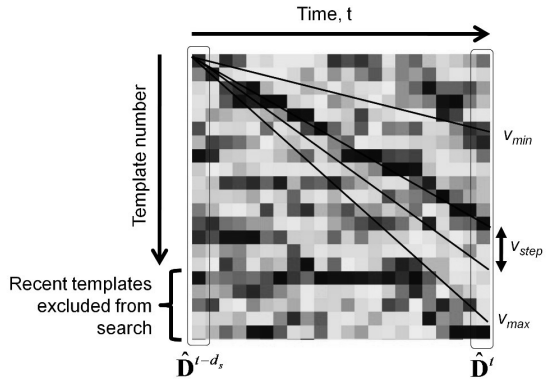

Figure 3: Searching for coherent matching sequences within the space of recent image difference vectors. Darker shading = smaller image difference = stronger match.

After all the trajectory scores have been evaluated, the minimum scoring (i.e. best matching) trajectory for each template is placed in vector $\mathbf{S}$. If the minimum scoring trajectory within a sliding window of range R_{window} is a factor of μ smaller than any of the trajectory scores outside the window, that trajectory is deemed to be a match (Fig. 4). μ is the parameter we vary to generate the precision-recall curves shown in the results section.

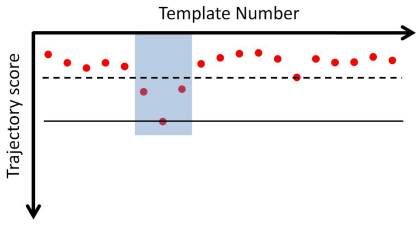

Figure 4: A trajectory within the sliding window is deemed a match if its score is significantly lower than the lowest trajectory score outside the sliding window.

4 Experimental Setup

In this section we describe the datasets used, ground truth measures, image pre-processing and parameter values.

4.1 Mountain Bike Dataset

The first dataset consisted of two videos of downhill mountain biking along a trail in the United States of America (Fig. 5). The videos were sourced from YouTube - http://www.youtube.com/watch?v=D4iU-EOJYK8 and http://www.youtube.com/watch?v=pTqh0ju5Z2o, using the default 360p versions. The first video was from a GoPro HD HERO mounted forward facing on the side of the rider's helmet, the second was from a GoPro HD HERO mounted on the bike handlebars. The only pre-processing performed was to crop the two videos to approximately the same field of view and joining the two videos together to form one 8734 frame, 640×480 resolution RGB video file (the actual "true resolution" of these videos is much lower than 640×480).

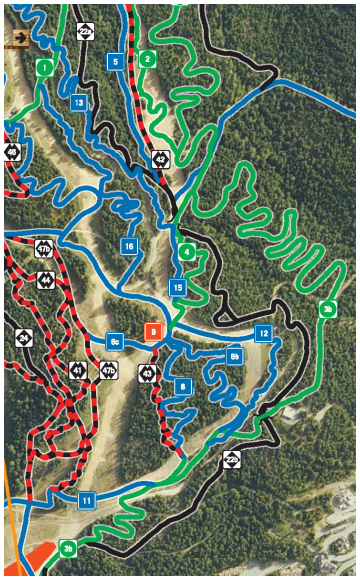

Figure 5: Aerial shot of the mountain bike path A-Line at Whistler which approximately corresponds to the path taken. Much of the track is in thick vegetation with highly varied lighting conditions. Image is sourced from a small section of the map available from http://www.whistlerbike.com. A-Line is the solid black track.

The quality of the video was quite low due to the combination of high compression ratio and high level of motion in the scenes. The field of view and source frame rate are unknown and not required for this work. The exact dataset gathering time of day and year are unknown, but appear to be from different times of day, but probably sometime between Spring and early Autumn, as there is no snow on the ground. A ground truth measure was obtained by manually labelling keyframe matches between the two datasets, and interpolating between these keyframes to achieve a continous ground truth measure.

4.2 Image Pre-Processing

The cropped full resolution images were changed to grayscale and down-sampled by a factor of 16 – bringing a 640x480 video down to 40 x 30 pixels (Figure 6). The OpenCV pixel area resampling technique was used to avoid Moire interference patterns. After down sampling, the image was divided into smaller square patch regions of side length P, which were then patch normalized. Patch normalization is a process which has been used by many roboticists to somewhat reduce the effects of illumination invariance [7, 28].

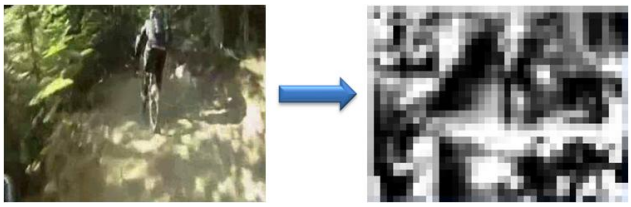

Figure 6 – Original color images were grayscaled, downsampled to a low resolution image, and patch normalized.

4.3 FAB-MAP Comparison

To provide a measure of how a feature-based mapping technique would perform on the dataset, we applied the open source version of FAB-MAP, openFABMAP, available at http://code.google.com/p/openfabmap/. To maximize recognition, we used the FAB-MAP v1.0 without fast bailout, and trained it on the first of the two runs down the mountain. We then ran it on the entire combined video (run1 + run2). A number of options were tried to maximize FAB-MAP performance, such as frame-subsampling. "Optimal" performance for this dataset was obtained by training on a decimated by 10 video. Apart from frame-subsampling, the default parameters provided in the openFABMAP (Version 1.01) implementation were used. We provide the source video links in the hope that others will also apply their feature-based algorithms.

4.4 Parameter Values

The table below provides the values for the various system parameters, which have been used on a number of datasets successfully.

Table 1: Parameter Values

Parameter	Value	Description
R_x, R_y	40, 30	Reduced image size, mountain bike dataset
R_{window}	10 templates	Local template neighborhood range
R_{recent}	20 templates	Recent template range

d_s	400 frames	Trajectory length in number of frames
d_{thresh}	50 frames	Maximum frame separation for match to be labelled correct
V_{av}	0.5 templates / frame	Average templates learned per frame
V_{min}	$0.6V_{av}$	Minimum trajectory speed
V_{max}	$1.48V_{av}$	Maximum trajectory speed
V_{step}	$0.04V_{av}$	Trajectory speed step-size
P	10 pixels	Patch normalization patch side length
μ	$1.0 \leq \mu \leq 2.0$	Trajectory uniqueness parameter

5 Results

In this section, we present the place recognition performance of the algorithm on the second run down the mountain, having already processed the first run. We generate precision-recall curves for both the algorithm and FAB-MAP 1.0 to facilitate comparison [6]. We also examine images from two illustrative matched subroutes. A video of the results is attached with this paper and also available at https://wiki.qut.edu.au/download/attachments/104094381/milford_acra.mp4

Precision-recall curves were generated by varying the trajectory uniqueness parameter μ. Each frame was classified as true positive (TP), false positive (FP), true negative (TN) or false negative (FN). In cases where the trajectory uniqueness score of the best matching subroute was larger than the trajectory uniqueness parameter μ (deemed a match), the frame was classified as either a true positive or false positive, depending on whether the frame discrepancy s was smaller or larger than d_{thresh}, respectively:

$$s = abs(f_r - f_{gt}) \quad (5)$$

where f_r is the recognized frame and f_{gt} is the ground truth frame. If the trajectory uniqueness score was below the trajectory uniqueness parameter μ, the frame was classified as a false negative. The d_{thresh} value of 50 frames corresponds to 1.7 seconds of real-time, which is a distance of 18 metres at 40 km/hr. For the mountain bike dataset, true negatives were not possible during the second run down the mountain.

5.1 Mountain Bike Dataset

Figure 7 shows the precision recall graph for the mountain bike dataset. At a precision level of 100%, the highest recall rate is 41%, after which the precision rate drops slowly. At the highest level of recall tested – 99%, the precision was 88%. In comparison, FAB-MAP 1.0 clearly struggles with the dataset, and peaks at about 11% precision around 5% recall. The FAB-MAP curve is not arbitrarily truncated – the rightmost datapoint corresponds to a matching threshold of zero, meaning any non-zero matching score counts as a "recalled" frame, whether a true positive or false positive.

Figures 8 and 9 show matching matrices for FAB-MAP and the proposed algorithm, respectively, for the second run down the mountain and truncated vertically to show only the region of the matching matrix that should have, ideally, a strong matching increasing diagonal. The darker the shading, the stronger the frame match. FAB-MAP is able to produce relatively few strong matches, and many of them are clearly false positives.

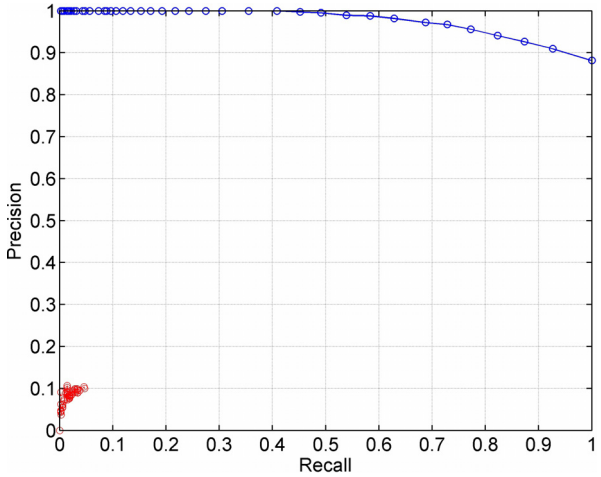

Figure 7 – Precision-recall graph for the mountain bike dataset, using local sequence matching (top solid line) and FAB-MAP 1.0 (bottom dashed line).

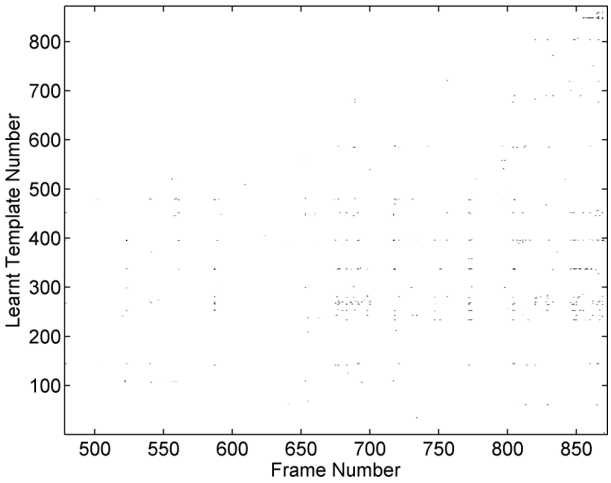

Figure 8 – Frame matches for FAB-MAP, with very few matches and no clear matching diagonal. Note the x-axes for both matching matrix plots show the processed frame number, which is a factor of 10 and 4 smaller than the original video frame number for FAB-MAP and SeqSLAM, respectively.

Figure 10 shows the corresponding images from a trajectory segment which was matched with a very high matching score. The images in the left column were generated by sampling at 5 even intervals over the entire 400 frame local trajectory segment in run 2, while the images in the right column are the corresponding images from the matching trajectory in the first run. Each pair of images represent the same place, despite some significant variations in lighting and the actual environment. For example, compare the shadowing in the 3rd and 4th row. Figure 11 shows the corresponding images for a matching trajectory segment at a different location.

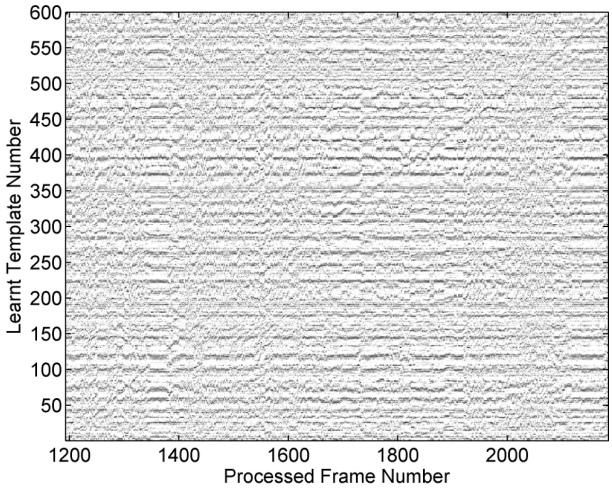

Figure 9 – Frame matches for the local sequence matching algorithm, with a clearly identifiable matching diagonal. The colour scale for this graph has been multiplied by a factor of 2 to enhance visibility. The graph corresponds approximately to the schematic graph in Figure 3.

Figure 10 – Corresponding frames from a matched 400 frame long trajectory sequence in the mountain bike dataset.

The allowable speed variation (expressed as a ratio) between the two downhill runs was 0.60 to 1.48, as shown by the dotted lines in Figure 12. As can be seen by the rarity with which the calculated speed ratio hits either of the two limits, the majority of the second downhill run was at least 60% of the speed of the first run and less than 48% faster. Note that the figure shows the calculated speed ratio for every sequence match, not just those that were matched with a certain confidence.

Figure 11 – Corresponding frames from a second matched 400 frame long trajectory sequence in the mountain bike dataset.

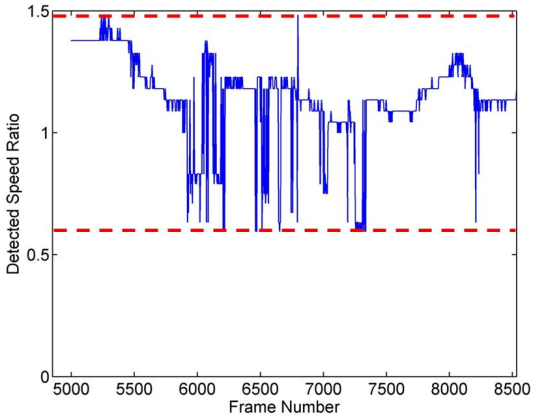

Figure 12 – Dataset speed ratio as calculated using the matching information. The dotted lines show the maximum speed ratio between traverses that the algorithm will match for the chosen experimental parameters.

5.2 Preliminary Day-Night Sun-Rain Results

We have applied the same techniques to a dataset obtained from a car driving two passes of an 8 km route through the Brisbane suburb Alderley. The first pass was obtained in the morning of a bright clear day. The second pass was obtained in the middle of the night during a thunderstorm with torrential rain. A Panasonic TZ-7 camera was positioned in a forward facing position inside the

windshield. As can be seen from Figure 13 (and confirmed independently), each of the corresponding images from the matching trajectory segment does match its image pair from the other pass. At this stage the maximal performance has been a recall rate of 35% at 100% precision. Work on larger datasets and geometric ground truth evaluation is ongoing.

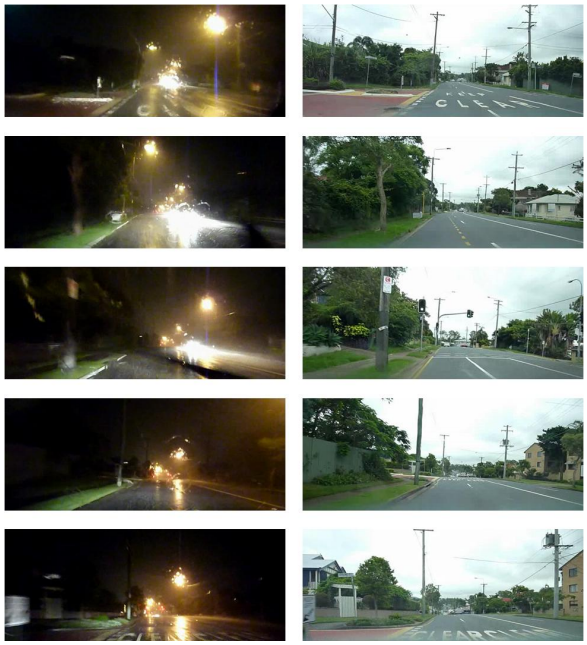

Figure 13 – Corresponding frames from matching trajectory segments (320 frame segment) from the clear daytime and rain nighttime Alderley datasets.

6 Conclusions and Future Work

In this paper we have presented an algorithm for localizing along a route even with significant perceptual change by matching image sequences rather than individual frames. By reducing the image comparison problem to one of only finding the best matching image within local sections of a route, we remove the requirement for a visual processing front-end that can reliably pick matches correctly in a global sense. Consequently we are able to use quite low quality visual information, such as the 1200 pixel images used in this work. Furthermore, the system works well even when there are no features that can be reliably detected across different times and conditions, rendering it vastly superior in such situations to feature-based techniques such as FAB-MAP 1.0, which we also compared. FAB-MAP was compared under generally favourable circumstances – it was trained on part of the actual test data, and was run on the full 640×480 pixel resolution images, while the sequence matching algorithm was run on 40×30 pixel resolution images, and was not pre-trained. However, the quality of the videos was far from sharp when 'pixel-peeping', which is perhap a significant cause of the poor FAB-MAP performance – the "effective resolution" of the videos was far less than 640×480 pixels. We also note that the proposed algorithms are currently subject to two constraining assumptions that we address in the following future work sections.

6.1 Locally Repeatable Odometry

The current implementation assumes that repeated traverses of a route are performed at approximately the same rate (in terms of video frames per distance travelled) – see Figure 12. In the day-night dataset work, we have implemented a simple no-motion detector which deals with the camera pausing in its motion. However, to produce a generally applicable solution with no velocity constraints at all, we are looking at using *locally repeatable odometry* to generate speed measures which are consistent in any particular part of the environment, but not necessarily globally consistent. The speed measure is then used to inform the sequence matching algorithm, enabling it to deal with large discrepancies in speed between repeated traverses of a route. The goal of achieving local repeatability is orthogonal to the usual aim of achieving consistent and globally accurate odometry and avoids several of the difficulties conventional feature-based visual odometry systems face in adverse environmental conditions.

6.2 Pose Invariance

In this paper, there is no allowance for offsets or changes in viewing perspective between repeated traverses of a route. In our past work [27], we have used simple offset matching to achieve some level of invariance to camera pose. We will continue to examine semi-pose-invariant image comparison techniques which do not explicitly rely on feature detection.

Acknowledgement

We thank Arren Glover for his help in implementing OpenFABMAP for this paper.

References

[1] K. Konolige and M. Agrawal, "FrameSLAM: From Bundle Adjustment to Real-Time Visual Mapping," *IEEE Transactions on Robotics*, vol. 24, pp. 1066-1077, 2008.

[2] M. Cummins and P. Newman, "Highly Scalable appearance-only SLAM - FAB-MAP 2.0," presented at Robotics Science and Systems, Seattle, 2009.

[3] A. J. Davison, I. D. Reid, N. D. Molton, and O. Stasse, "MonoSLAM: Real-Time Single Camera SLAM," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 29, pp. 1052-1067, 2007.

[4] H. Strasdat, J. M. Montiel, and A. J. Davison, "Scale Drift-Aware Large Scale Monocular SLAM," presented at Robotics Science and Systems, Zaragoza, Spain, 2010.

[5] C. Valgren and A. Lilienthal, "Sift, surf, and seasons: Long-term outdoor localization using local features," in *Proc. of 3rd European Conference on Mobile Robots*. Freiburg, Germany, 2007.

[6] A. J. Glover, W. P. Maddern, M. J. Milford, and G. F. Wyeth, "FAB-MAP + RatSLAM: Appearance-based SLAM for Multiple Times of Day," presented at International Conference on Robotics and Automation, Anchorage, United States, 2010.

[7] M. Milford and G. Wyeth, "Persistent Navigation and Mapping using a Biologically Inspired SLAM System," *International Journal of Robotics Research*, vol. 29, pp. 1131-1153, 2010.

[8] M. Milford and G. Wyeth, "Hybrid robot control and SLAM for persistent navigation and mapping," *Robotics and Autonomous Systems*, vol. 58, pp. 1096-1104, 2010.

[9] M. Milford and G. Wyeth, "Mapping a Suburb with a Single Camera using a Biologically Inspired SLAM System," *IEEE Transactions on Robotics*, vol. 24, pp. 1038-1053, 2008.

[10] D. G. Lowe, "Object recognition from local scale-invariant features," presented at International Conference on Computer Vision, Kerkyra, Greence, 1999.

[11] H. Bay, T. Tuytelaars, and L. Van Gool, "SURF: Speeded Up Robust Features," in *Computer Vision – ECCV 2006*, 2006, pp. 404-417.

[12] M. Cummins and P. Newman, "Highly scalable appearance-only SLAM - FAB-MAP 2.0," presented at Robotics: Science and Systems, Seattle, United States, 2009.

[13] M. Milford and G. Wyeth, "Single Camera Vision-Only SLAM on a Suburban Road Network," presented at International Conference on Robotics and Automation, Pasadena, United States, 2008.

[14] H. Andreasson, T. Duckett, and A. Lilienthal, "A Minimalistic Approach to Appearance-Based Visual SLAM," *IEEE Transactions on Robotics*, vol. 24, pp. 1-11, 2008.

[15] L. M. Paz, P. Pinies, J. D. Tardos, and J. Neira, "Large-Scale 6-DOF SLAM With Stereo-in-Hand," *IEEE Transactions on Robotics*, vol. 24, pp. 946-957, 2008.

[16] G. Dissanayake, P. M. Newman, S. Clark, H. Durrant-Whyte, and M. Csorba, "A solution to the simultaneous localisation and map building (SLAM) problem," *IEEE Transactions on Robotics and Automation*, vol. 17, pp. 229-241, 2001.

[17] L. Clemente, A. J. Davison, I. D. Reid, J. Neira, and J. D. Tardos, "Mapping large loops with a single hand-held camera," presented at Robotics: Science and Systems, Atlanta, United States, 2007.

[18] G. Sibley, C. Mei, I. Reid, and P. Newman, "Vast-scale Outdoor Navigation Using Adaptive Relative Bundle Adjustment," *International Journal of Robotics Research*, vol. 29, pp. 958-980, 2010.

[19] P. Biber and T. Duckett, "Dynamic Maps for Long-Term Operation of Mobile Service Robots," presented at Robotics: Science and Systems, Cambridge, United States, 2005.

[20] S. Nuske, J. Robert, and G. Wyeth, "Robust outdoor visual localization using a three-dimensional-edge map," *Journal of Field Robotics*, vol. 26, pp. 728-756, 2009.

[21] A. E. Abdel-Hakim and A. A. Farag, "CSIFT: A SIFT descriptor with color invariant characteristics," presented at IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2006.

[22] K. Van de Sande, T. Gevers, and C. Snoek, "Evaluation of color descriptors for object and scene recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 32, pp. 1582-1596, 2008.

[23] J. A. Castellanos, J. Neira, and J. D. Tardós, "Multisensor fusion for simultaneous localization and map building," *IEEE Transactions on Robotics and Automation*, vol. 17, pp. 908-914, 2002.

[24] J. S. Gutmann and K. Konolige, "Incremental mapping of large cyclic environments," presented at Conference on Intelligent Robots and Applications, 1999.

[25] E. Shechtman and M. Irani, "Matching local self-similarities across images and videos," presented at IEEE Conference on Computer Vision, Minneapolis, United States, 2007.

[26] M. J. Milford, G. Wyeth, and D. Prasser, "RatSLAM: A Hippocampal Model for Simultaneous Localization and Mapping," presented at IEEE International Conference on Robotics and Automation, New Orleans, USA, 2004.

[27] M. Milford, F. Schill, P. Corke, R. Mahony, and G. Wyeth, "Aerial SLAM with a Single Camera Using Visual Expectation," presented at International Conference on Robotics and Automation, Shanghai, China, 2011.

[28] A. M. Zhang, "Robust Appearance Based Visual Route Following in Large Scale Outdoor Environments," presented at Australasian Conference on Robotics and Automation, Brisbane, 2007.